UNIVERSITETET
I OSLO

Data storage considerations for HTS platforms

George Magklaras -- node manager

http://www.no.embnet.org

http://www.biotek.uio.no

admin@embnet.uio.no

Overview:

- The need for data storage
- Volume dimensioning – the right size
- The right filesystem for the right job
- How to benchmark filesystems
- The network layer - FcOE as storage barrier braker
- Credits
- Questions

The presentation sales pitch:

- I am a biologist, give me a break, this is too technical.
  - Yes, but you will have to deal with it.
  - Instruments produce lots of data whether you are technical or not.
  - The number of instruments is surpassing the ability of departmental infrastructures to store data.
  - Your IT people would find this info useful.

The Norwegian EMBnet platform

HTS Device	**No. of runs Per year**	**Tier 1 Gbytes**	**Tier 2 Gbytes**	**Tier 3 Gbytes**	**Tier 4 Gbytes**	**Total *Tbytes***
Illumina	100	9728	100	300	400	990
454	100	200	50	25	75	27
SOLiD	100	6144	100	100	200	80

Volume dimensioning (2)

- Tier 1: raw unprocessed data as they come out from the instrument (mostly images). For most HTS devices, Tier 1 data will generate several Tbytes per run (several thousands of Gigabytes), especially as the instrument's ability to become more precise gets better with time (firmware or device upgrades).
- Tier 2: Initial processing data stage: including base (or colour) calls, intensities and first pass quality scores. These data are currently in the order of several tenths of Gigabytes to 300 Gigabytes per run maximum for certain types of sequencers.
- Tier 3: Includes aligned and analyzed data (alignments of all the reads to a reference or de-novo assembly, if required). This can be at least as big as the initial processing stage (Tier 2), since the initial reads themselves have to be preserved as part of the alignment output. At the end of each successful processing step, the raw data of Tier 1 are removed.

Volume Dimensioning (3)

- Tier 4: The final fourth tier includes data that should be backed up off site, in order to provide disaster recovery, as well as a long term archive. This includes a mirror of Tier 2 and 3 plus the archive requirements. It is not financially feasible or technically practical to off-site backup Tier 1 data, at least not for every run, as the volume of data is huge. There is some data redundancy between tiers 2 and 3, as in theory one could resort Tier 3 reads according to the alignment output and then discard Tier 2 data. However, this might not be feasible/desirable in all analysis scenarios and thus we assume it is good practice to backup and archive both Tier 2 and Tier 3 data.

Volume dimensioning (4)

$$\mathbf{Tier1}_{store}\mathbf{=\Sigma(N}_{hts}\ \mathbf{x\ G}_{bpr}\ \mathbf{+\ (N}_{hts}\ \mathbf{x\ G}_{bpr}\mathbf{)/4)\ (x\ Nruns)}$$

N_{hts}=*number of per type HTS devices*, G_{bpr}=*Gigabytes per run*

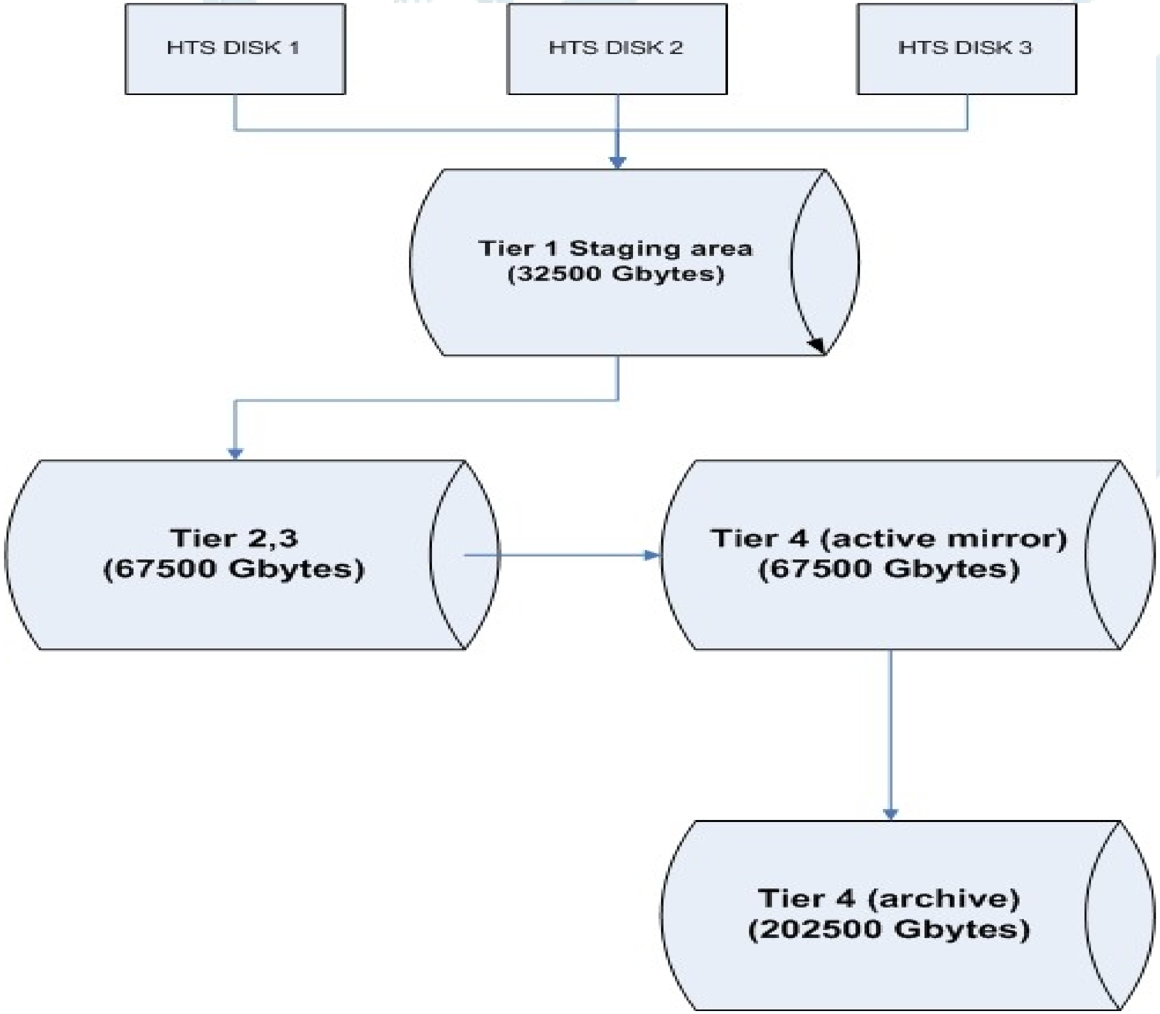
$$\boldsymbol{Tier2,3}_{store}\boldsymbol{=\Sigma(N}_{runs}\ \boldsymbol{x\ G}_{analysis}\ \boldsymbol{+\ (N}_{runs}\ \boldsymbol{x\ G}_{analysis}\boldsymbol{)/3)}$$

N_{runs}=*expected number of runs per year,*
$G_{analysis}$=*Gigabytes per run for Tiers 2 and 3 (Table 1)*

$$\boldsymbol{Tier4}_{store}\boldsymbol{=Tier2,3}_{store}\ \boldsymbol{+\ R}_{period}\ \boldsymbol{x\ Tier2,3}_{store}$$

R_{period}= *number of years to keep the data*

Volume Dimensioning (5)



- 2 x Illumina
- 2 x 454
- 1 x SOLiD
- 3 year data retention period

Conclusion:

Next Gen Sequencing is a classic example of data intensive computing [1].

Tier facilitate compartmentalization, because the number and range of tasks are different.

Filesystems galore

- A filesystem is a key component of the Operating System that dictates how the files are stored and accessed.
- **Commonly used disk filesystems**: ext3/4 (Linux) [2,3], NTFS (Windows) [4], HFS+ (MACOSX)[5], ZFS (Sun)[6]*****.
- **Shared/clustered/SAN filesystems**: GFS (RedHat)[7],XSAN (Apple)[8]
- **Distributed File Systems (Network File Systems)**: NFS(9), CIFS/SMB(10)
- **Distributed parallel fault-tolerant file systems**: GPFS (IBM) [11], XtreemeFS[12], OneFS (Isilon) [13], PanFS(Panasas)[14], Lustre [15]

Filesystem requirements:

- How to choose then?
- Next Gen Sequencing (NGS) filesystems need to be:
  - **Scalable in size**: Ext3 with max. volume size of 16 TiB would not fit the bill.
  - **Scalable in the number of IOPS for read/writes/nested directory access:** (NTFS would not scale that well here).
  - **Allow concurrent access**: Raw data accessed by hundreds/thousands of compute nodes.
  - **Have file redundancy/replication features**: Have you ever measured restore times for multiple TiB volumes? What if you want to replicate part of the data set across regions/countries/continents for your colleagues?

Filesystem requirements:

- Ideally they should offer transparent disk encryption features: sensitive sequences/hospital environments...

- All these criteria point to distributed parallel fault-tolerant filesystems:
  - Distributed: Data replication issues
  - Parallel: Raise the IOPS gauge and facilitate concurrent access.
- Tier 1: Disks FS (ext4 + CTDB [16]) : To Facilitate cross-platform access
- Tiers 2,3: Lustre
- Tier 4: Lustre + other candidates

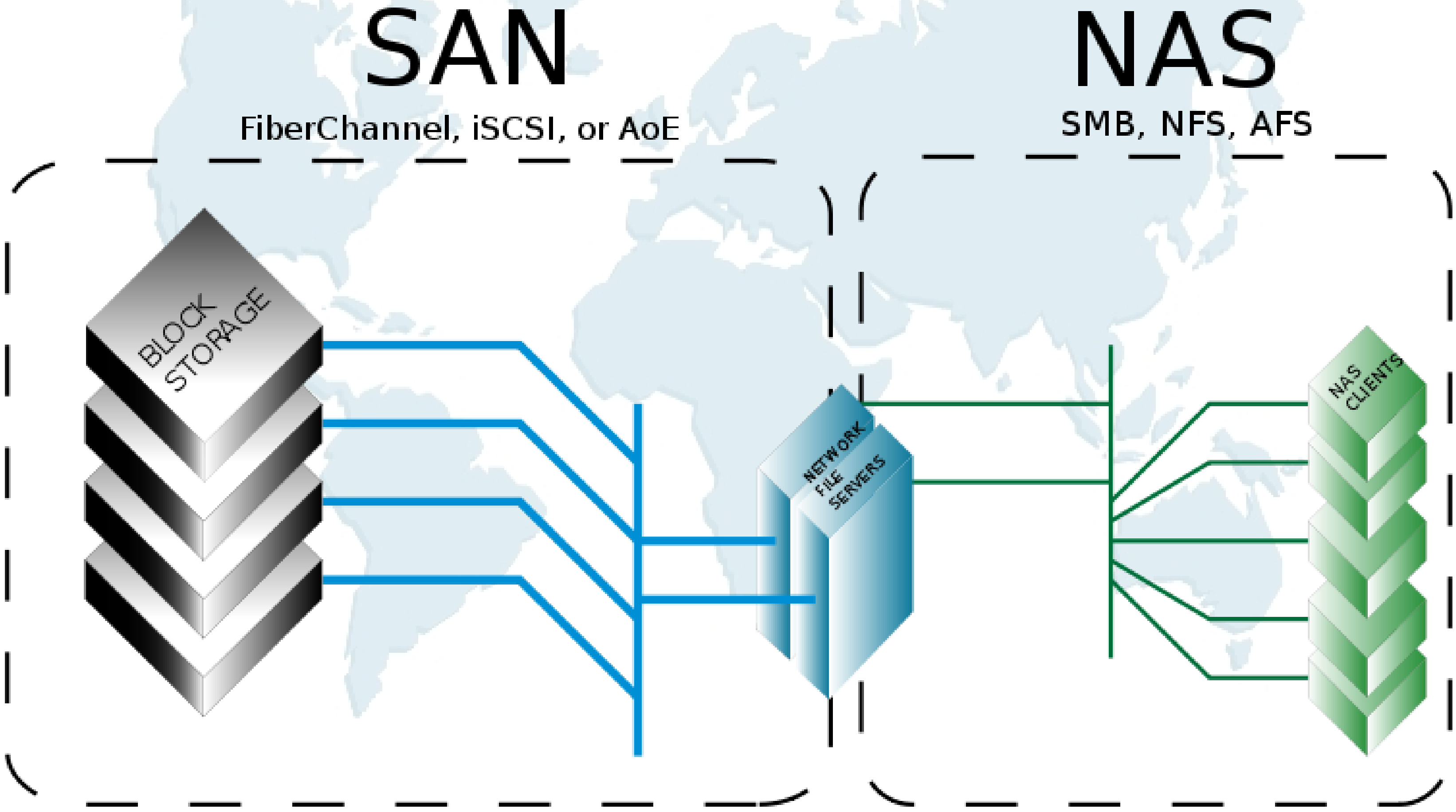
Filesystem benchmarks

FS	Random I/O (MB/s)	Seq. I/O MB/S	IOPS/seek	Recovery time
ext3	40/20	1/132	2300	2 days
ext4	76/48	90/72	4800	2 days
GPFS	114/290	187/320	6700	3 days
Lustre	178/190	380/336	5800	5 days

The data network and storage

- **DAS** versus **NAS** versus **SAN:**
  - **Directly Attached Storage (DAS)**: SATA/SAS, simple, effective for single/dual host access for capacities of up to 20/30 Tbytes. The cheapest, but not the least scalable in terms of storage capacity and simultaneous access.
  - **Network Attached Storage(NAS)**: A combination of a TCP/IP based protocol and a filesystem (NTFS over SMB/CIFS, NFS over ext3/4). Medium price and scalability.
  - **Storage Area Network(SAN)**: Block storage over a proprietary (Fiber Channel or off-the shelf protocol (iSCSI, AoE). Expensive, fast.

The data network and storage (2)

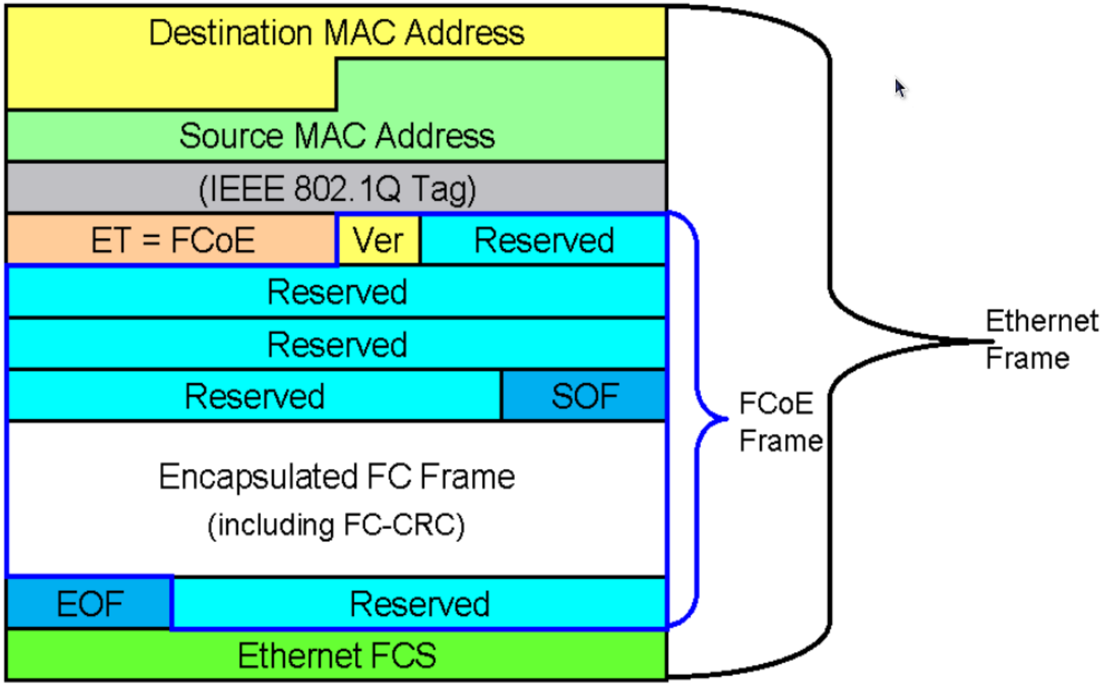


The data network and storage (3)

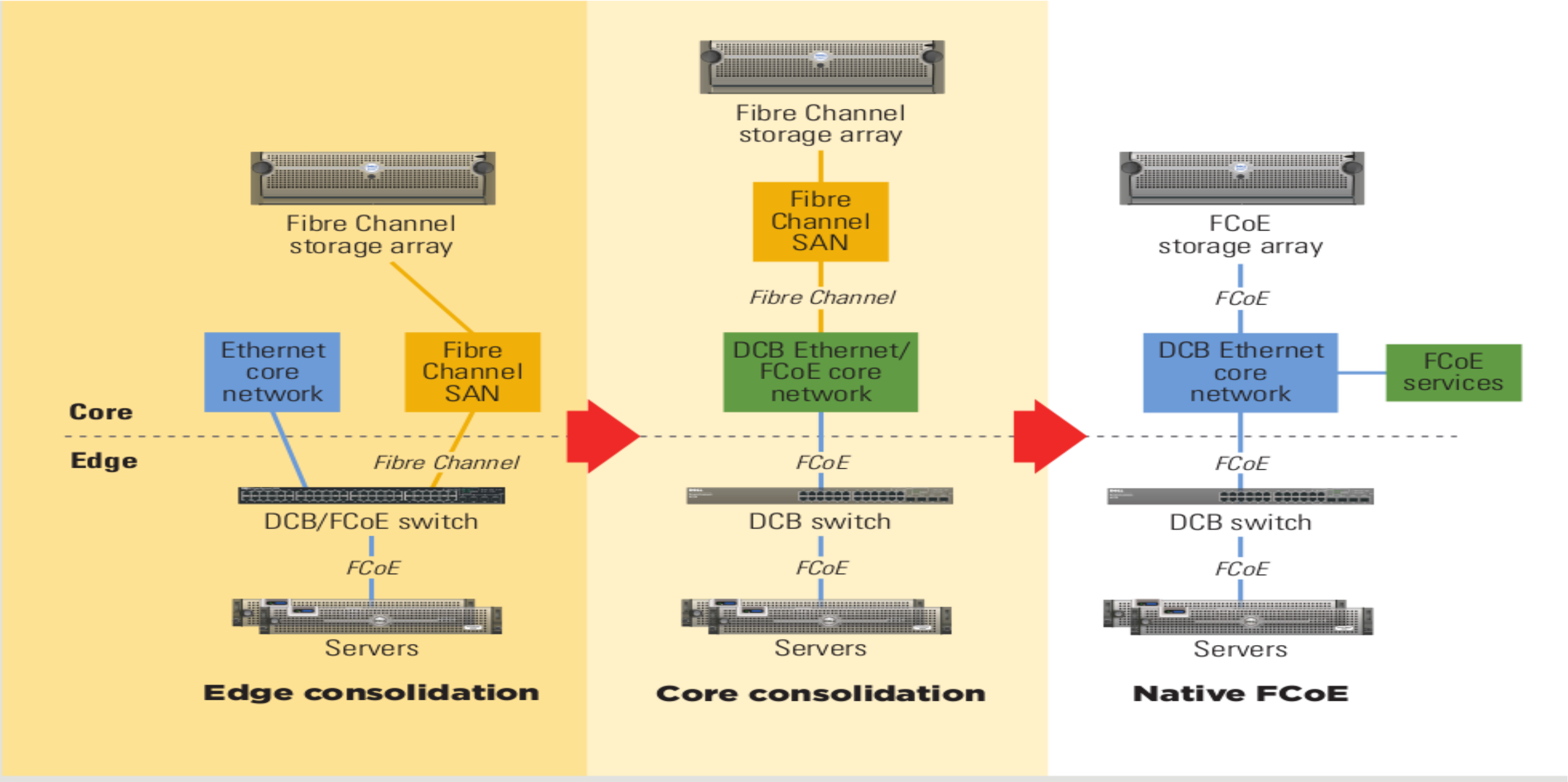
Questions for your IT architect/system administrator(s):

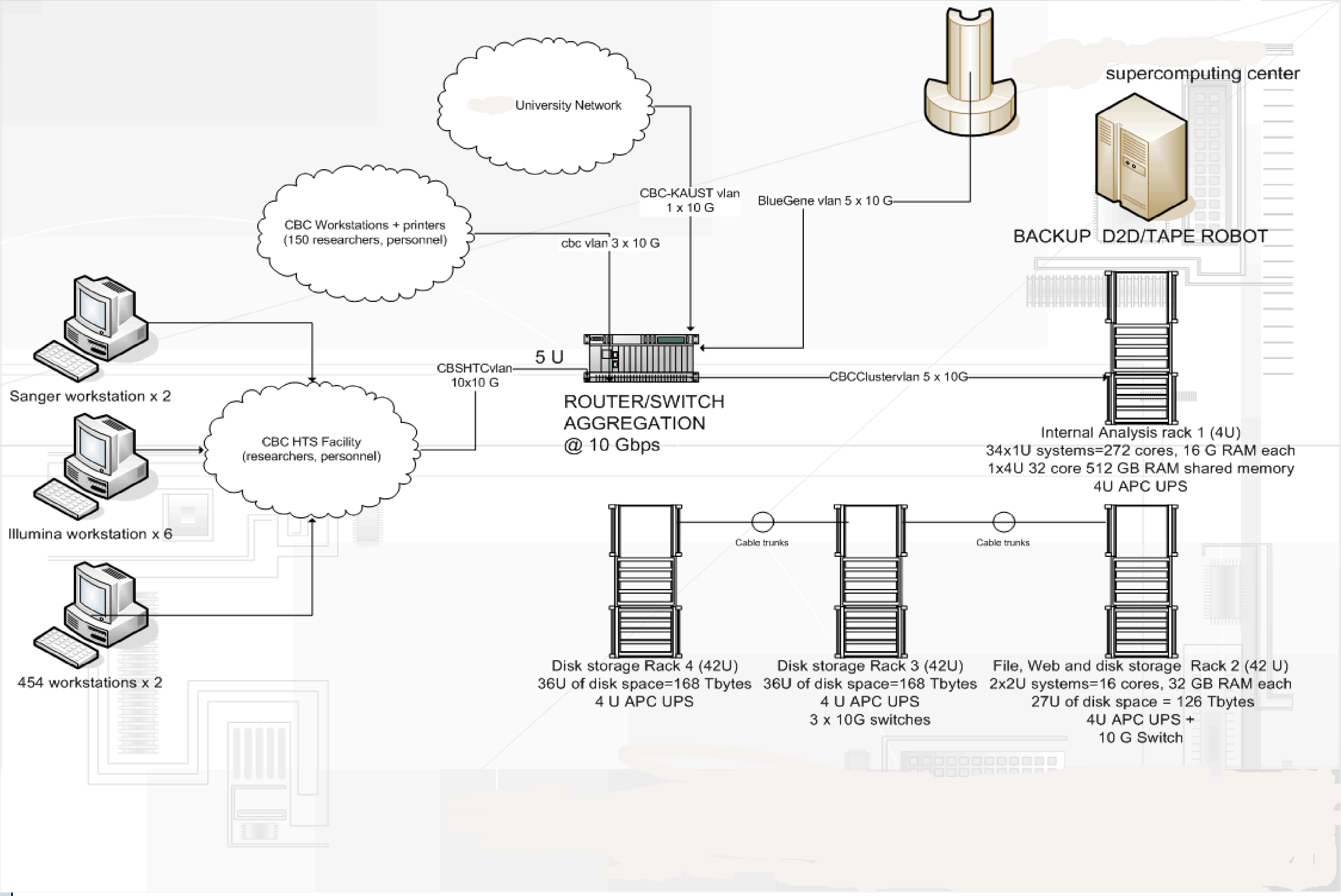
- Can you afford the pure Fiber Channel solutions today?
- How many storage interconnects you have (GigE, FC, Infiniband).
- Would it not be nice to have a smaller number of storage interconnects (consolidation)?
- If you have already chosen iSCSI, have you wondered how much is the overhead of encapsulating block protocols over TCP/IP?

The data network and storage - FCoE (4)



The data network and storage - FCOE(5)





University Network

CBC Workstations + printers
(150 researchers, personnel)

CBC-KAUST vlan
1 x 10 G

BlueGene vlan 5 x 10 G

cbc vlan 3 x 10 G

supercomputing center

BACKUP D2D/TAPE ROBOT

Sanger workstation x 2

Illumina workstation x 6

454 workstations x 2

CBC HTS Facility
(researchers, personnel)

CBSHTCvlan
10x10 G

5 U

ROUTER/SWITCH
AGGREGATION
@ 10 Gbps

CBCClustervlan 5 x 10G

Internal Analysis rack 1 (4U)
34x1U systems=272 cores, 16 G RAM each
1x4U 32 core 512 GB RAM shared memory
4U APC UPS

Cable trunks

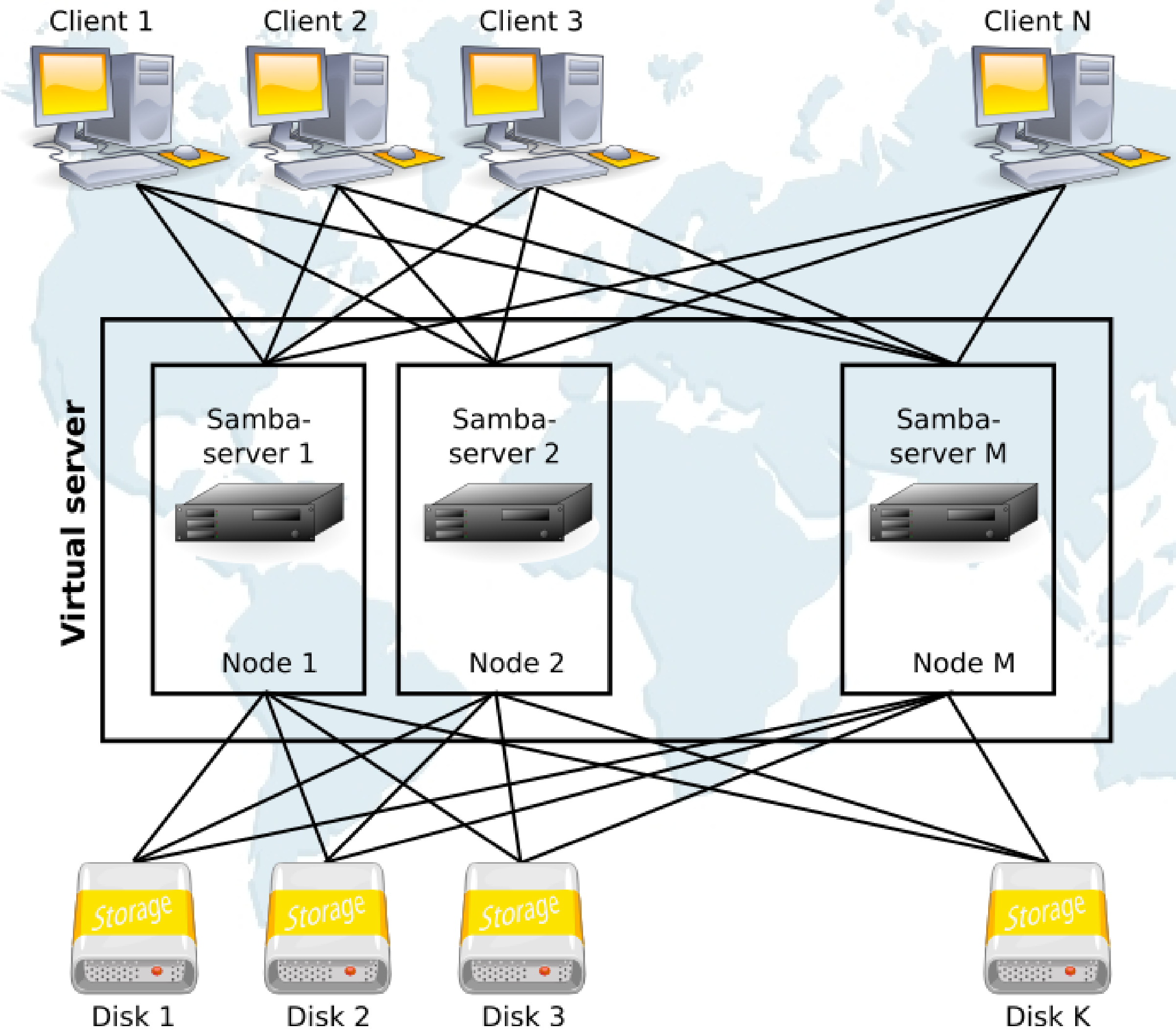
Cable trunks

Disk storage Rack 4 (42U)
36U of disk space=168 Tbytes
4 U APC UPS

Disk storage Rack 3 (42U)
36U of disk space=168 Tbytes
4 U APC UPS
3 x 10G switches

File, Web and disk storage Rack 2 (42 U)
2x2U systems=16 cores, 32 GB RAM each
27U of disk space = 126 Tbytes
4U APC UPS +
10 G Switch

Clustered Samba – Tier 1 entry



Bill of materials:

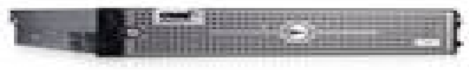
Cisco Nexus Switch 5000 series switches

Bill of materials (2):

QLE8152
Dual Port 10GbE Ethernet to PCIe Converged Network Adapter (CNA).
www.qlogic.com

Bill of materials (3):

Dell | EMC CX4-960
(8Gbit and 4Gbit FC/10 and
1Gbit iSCSI SAN Array)



Dell 1950, 64 Gbytes of
RAM/Qlogic CN cards (as
access/front end nodes),
8 cores.

Bill of materials (4):

Redhat Enterprise Linux 5.4 (has support for FCoE

Samba 3.3.x with CTDB (**not the one that comes with RHEL 5.4**)

Comments/questions:

email:

georgios@biotek.uio.no

Credits

UNIVERSITETET
I OSLO

EMBIO Styringsgruppen for
Molecular Life Sciences ved UiO

DELL

CISCO

redhat.

samba